

CLAIMS

1. A method for mining association rules in a database that is divided into multiple partitions associated with respective computing nodes, the method comprising:

transmitting messages among the nodes with respect to local support of an itemset in the respective partitions of the database;

responsive to the messages transmitted by a subset of the nodes, determining the itemset to be globally frequent in the database before the nodes outside the subset have transmitted the messages with respect to the local support of the itemset in their respective partitions; and

computing an association rule with respect to the itemset, responsive to having determined the itemset to be globally frequent.

2. A method according to claim 1, wherein transmitting the messages comprises conveying the messages over a communication network connecting the nodes one to another.

3. A method according to claim 2, wherein conveying the messages comprises broadcasting the messages.

4. A method according to claim 2, wherein conveying the messages comprises stacking a plurality of the messages together in a single data frame for transmission over the network.

5. A method according to claim 1, wherein transmitting the messages comprises computing a candidacy criterion at each of the nodes, for use in determining whether the itemset is globally frequent, and choosing the itemset

with respect to which one of the messages is to be transmitted responsive to the candidacy criterion.

6. A method according to claim 5, wherein computing the candidacy criterion comprises receiving one of the messages sent by another one of the nodes, and recomputing the candidacy criterion responsive to the local support conveyed by the received message, and

wherein choosing the itemset comprises deciding whether to transmit another one of the messages with respect to the itemset based on the recomputed criterion.

7. A method according to claim 5, wherein deciding whether to transmit another one of the messages comprises transmitting another one of the messages only until a conclusion is reached, responsive to the candidacy criterion, as to whether the itemset is globally frequent in the database.

8. A method according to claim 5, wherein transmitting the messages comprises terminating transmission of the messages when the candidacy criterion computed at every one of the nodes agrees as to whether the itemset is globally frequent.

9. A method according to claim 5, wherein the itemset is one of a plurality of itemsets in the database, and wherein computing the candidacy criterion comprises computing respective candidacy criteria for the plurality of the itemsets, and wherein choosing the itemsets comprises ranking the itemsets responsive to the respective candidacy criteria for transmission of the messages with respect thereto.

10. A method according to claim 9, wherein ranking the itemsets comprises determining a respective ranking for each of the nodes, and wherein transmitting the messages comprises selecting one of the nodes that is to transmit the messages, responsive to the respective ranking.

11. A method according to claim 10, wherein determining the respective ranking comprises updating the ranking as the messages are transmitted, and wherein selecting the one of the nodes comprises changing a selection of the one of the nodes that is to transmit the messages responsive to a change in the ranking.

12. A method according to claim 1, wherein computing the association rule comprises collecting the local support of the itemset from the nodes outside the subset, for use in computing the association rule applicable to the itemset, only after it is determined that the itemset is globally frequent.

13. A method according to claim 12, wherein the itemset is one of a plurality of itemsets in the database, and wherein collecting the local support comprises collecting the local support of the itemsets that were determined to be globally frequent, while ignoring the local support of the itemsets that were not determined to be globally frequent.

14. A method according to claim 12, wherein computing the association rule comprises assessing a confidence level of the rule responsive to the local support, and wherein collecting the local support comprises computing a confidence criterion at each of the nodes, for use in determining whether the confidence level is above a predetermined threshold, and choosing the itemset with

respect to which the local support is to be collected responsive to the confidence criterion.

15. A method according to claim 14, wherein computing the confidence criterion comprises receiving the local support sent by another one of the nodes, and recomputing the confidence criterion responsive to the received local support, and

wherein choosing the itemset comprises continuing to collect the local support until it is determined that the confidence level is above the predetermined threshold, based on the recomputed criterion.

16. A method according to claim 1, wherein the itemset is one of a plurality of itemsets in the database, each of the itemsets having a size, and wherein determining the itemset to be globally frequent comprises finding the itemsets of size k that are globally frequent, and wherein transmitting the messages comprises transmitting the messages with respect to the local support of the itemsets of size $k+1$ all of whose subsets are itemsets of size k that were found to be globally frequent.

17. A method for mining association rules in a database that is divided into multiple partitions associated with respective computing nodes, the partitions including at least first and second partitions respectively associated with at least first and second nodes among the computing nodes, the method comprising:

computing an initial candidacy criterion at each of the nodes, for use in determining whether an itemset is globally frequent in the database;

responsive to the candidacy criterion, transmitting a first message from the first node to the other nodes

conveying a local support of the itemset in the first partition;

upon receiving the message, recomputing the candidacy criterion at the second node responsive to the local support conveyed by the message;

transmitting, responsive to the recomputed candidacy criterion, a second message from the second node to the other nodes, conveying the local support of the itemset in the second partition; and

computing an association rule with respect to the itemset, responsive to the first and second messages.

18. A method according to claim 17, wherein the nodes further comprise a third node, and wherein the method comprises recomputing the candidacy criterion at the third node, responsive to the first and second messages, and determining at the third node that the itemset is globally frequent based on the recomputed criterion.

19. A method according to claim 18, wherein determining at the third node that the itemset is globally frequent comprises making a conclusive determination that the itemset is globally frequent before all the nodes have transmitted messages conveying the local support of the itemset in the respective partitions of the database.

20. A method according to claim 18, wherein computing the association rule comprises computing the rule responsive to having determined that the itemset is globally frequent based on the recomputed criteria.

21. A method according to claim 17, wherein computing the initial candidacy criteria comprises computing at each of the first and second nodes a local hypothesis as to whether the itemset is globally frequent, based on the

local support of the itemset in the first and second partitions, respectively, and

wherein recomputing the candidacy criterion comprises recomputing the local hypothesis and computing a global hypothesis as to whether the itemset is globally frequent, based on the local support conveyed in the first message, and

wherein transmitting the second message comprises deciding whether to transmit the second message responsive to the local and global hypotheses.

22. A method according to claim 21, wherein deciding whether to transmit the second message comprises deciding to transmit the second message only if the local and global hypotheses computed at the second node disagree as to whether the itemset is globally frequent.

23. A method according to claim 21, wherein the itemset is one of a plurality of itemsets in the database, and wherein computing and recomputing the candidacy criterion comprise computing and recomputing respective local and global hypotheses for the plurality of the itemsets, and wherein transmitting the first and second messages comprises choosing the itemset with respect to which the messages are to be transmitted responsive to the respective hypotheses.

24. A method according to claim 23, wherein choosing the itemset comprises ranking the itemsets responsive to a measure of disagreement between the local and global hypotheses with respect to the itemsets.

25. A method according to claim 23, wherein deciding whether to transmit the second message comprises, if none

of the local and global hypotheses disagree, transmitting a pass message.

26. Apparatus for mining association rules, comprising:

a plurality of storage devices, adapted to hold respective partitions of a database; and

a corresponding plurality of computing nodes, each node being associated with a respective one of the storage devices and coupled to communicate with the other nodes over a communication network, the nodes being adapted to transmit messages one to another with respect to local support of an itemset in the respective partitions of the database, and responsive to the messages transmitted by a subset of the nodes, to determine the itemset to be globally frequent in the database before the nodes outside the subset have transmitted the messages with respect to the local support of the itemset in their respective partitions, and to compute an association rule with respect to the itemset, responsive to having determined the itemset to be globally frequent.

27. Apparatus for mining association rules, comprising:

a plurality of storage devices, adapted to hold respective partitions of a database, including at least first and second storage devices holding respective first and second partitions of the database; and

a corresponding plurality of computing nodes, each node being associated with a respective one of the storage devices, including at least first and second nodes respectively associated with the first and second storage devices, the nodes being coupled to communicate with one another over a communication network, each of

the nodes further being adapted to compute an initial candidacy criterion, for use in determining whether an itemset is globally frequent in the database, such that responsive to the candidacy criterion, the first node transmits a first message to the other nodes conveying a local support of the itemset in the first partition, and such that upon receiving the message, the second node recomputes the candidacy criterion responsive to the local support conveyed by the message and transmits, responsive to the recomputed candidacy criterion, a second message from the second node to the other nodes, conveying the local support of the itemset in the second partition, so that the nodes compute an association rule with respect to the itemset responsive to the first and second messages.

28. A computer software product, comprising a computer-readable medium in which program instructions are stored, which instructions, when read by computing nodes that are associated with respective storage devices holding respective partitions of a database and are coupled to communicate with one another over a communication network, cause the nodes to transmit messages one to another with respect to local support of an itemset in the respective partitions of the database, and responsive to the messages transmitted by a subset of the nodes, to determine the itemset to be globally frequent in the database before the nodes outside the subset have transmitted the messages with respect to the local support of the itemset in their respective partitions, and to compute an association rule with

respect to the itemset, responsive to having determined the itemset to be globally frequent.

29. A computer software product, comprising a computer-readable medium in which program instructions are stored, which instructions, when read by computing nodes that are associated with respective partitions of a database, including at least first and second nodes respectively associated with first and second partitions of the database, and which are coupled to communicate with one another over a communication network, cause the nodes to compute an initial candidacy criterion, for use in determining whether an itemset is globally frequent in the database, and responsive the candidacy criterion, cause the first node to transmit a first message to the other nodes conveying a local support of the itemset in the first partition, and cause the second node, upon receiving the message, to recompute the candidacy criterion responsive to the local support conveyed by the message and to transmit, responsive to the recomputed candidacy criterion, a second message to the other nodes, conveying the local support of the itemset in the second partition, and cause the nodes to compute an association rule with respect to the itemset responsive to the first and second messages.

30. A method for processing items in a database that is divided into multiple partitions associated with respective computing nodes, the method comprising:

transmitting messages among the nodes conveying local information regarding an itemset in the respective partitions of the database;

responsive to the messages transmitted by a subset of the nodes, determining the itemset to be globally significant with respect to a decision to be made in reference to the database before the nodes outside the subset have transmitted the messages with respect to the local information regarding the itemset in their respective partitions; and

making the decision with respect to the itemset, responsive to having determined the itemset to be globally significant.

31. A method for reaching a decision regarding items in a database that is divided into multiple partitions associated with respective computing nodes, the partitions including at least first and second partitions respectively associated with at least first and second nodes among the computing nodes, the method comprising:

computing an initial candidacy criterion at each of the nodes, for use in selecting an itemset in the database of potential significance to the decision;

responsive to the candidacy criterion, transmitting a first message from the first node to the other nodes conveying local information regarding the itemset in the first partition;

upon receiving the message, recomputing the candidacy criterion at the second node responsive to the local information conveyed by the message;

transmitting, responsive to the recomputed candidacy criterion, a second message from the second node to the other nodes, conveying the local information regarding the itemset in the second partition; and

making the decision with respect to the itemset,
responsive to the first and second messages.

32. A method according to claim 31, wherein computing
the initial candidacy criterion comprises determining the
candidacy criterion based on a target function selected
responsive to the decision that is to be made.